

Claims

1. A method of detecting a chromosomal abnormality based on artificial intelligence, the method comprising:

- a) obtaining sequence information using extracted nucleic acids from a biological sample;
- b) aligning the obtained sequence information (reads) with a reference genome database;
- c) generating vectorized data using nucleic acid fragments based on the aligned sequence information (reads); and
- d) inputting the generated vectorized data into a trained artificial intelligence model, analyzing the data, and comparing the resulting value with a cut-off value to determine whether or not a chromosomal abnormality is present,

wherein the vectorized data of step (c) is a grand canyon plot (GC plot) or a genomic castle wall plot (GCW plot),

wherein the GC plot is generated by calculating a distribution of aligned nucleic acid fragments in each chromosome bin based on the distance between the nucleic acid fragments using a process including the following steps:

- i) dividing chromosomes into predetermined bins;
- ii) calculating the distance between nucleic acid fragments (fragment distance, FD) aligned in each bin;
- iii) determining a representative distance between fragments (RepFD) of each bin based on the fragment distance calculated in each bin;
- iv) normalizing RepFD by dividing the representative distance between fragments (RepFD) calculated in step iii) by a representative total nucleic acid fragment distance; and
- v) generating the GC plot with the order of respective bins on an X-axis and the normalized value calculated in step iv) on a Y-axis;

wherein the GCW plot is created using a process including the following steps:

- 1) calculating the distance between the aligned nucleic acid fragments (FD) for each chromosome;
- 2) determining a representative distance between fragments for each bin based on the

distance calculated in step i);

3) standardizing the representative distance between fragments in each chromosome determined in step ii);

4) selecting a portion of chromosomes excluding chromosomes for which presence of aneuploidy is to be determined as control chromosomes; and

5) generating the GCW plot with the control chromosomes and the chromosomes for which presence of aneuploidy is to be determined sequentially and alternately aligned on the X axis and the standardized calculated values of the respective chromosomes on the Y axis,

wherein the sequentially and alternately alignment is aligning ([calculated value of step ii) of the control chromosome], [calculated value of step ii) of the chromosome for which presence of aneuploidy is to be determined]) along the x-axis and

wherein the artificial intelligence model of step (d) is trained to distinguish between vectorized data of normal chromosomes and vectorized data of abnormal chromosomes.

2. The method according to claim 1, wherein step (a) comprises:

(a-i) obtaining nucleic acids from blood, semen, vaginal cells, hair, saliva, urine, oral cells, amniotic fluid containing placental cells or fetal cells, tissue cells, or a mixture thereof;

(a-ii) obtaining purified nucleic acids by removing proteins, fats, and other residues from the obtained nucleic acids using a salting-out method, a column chromatography method, or a bead method;

(a-iii) preparing a single-end sequencing or paired-end sequencing library for the purified nucleic acids or nucleic acids randomly fragmented by an enzymatic digestion, pulverization, or hydroshear method;

(a-iv) reacting the prepared library with a next-generation sequencer; and

(a-v) obtaining sequence information (reads) of the nucleic acids in the next-generation sequencer.

3. The method according to claim 1, wherein the representative FD (RepFD) comprises at least one selected from the group consisting of a sum, difference, product, mean, median,

quantile, minimum, maximum, variance, standard deviation, median absolute deviation, coefficient of variance of FD, a reciprocal thereof and a combination thereof.

4. The method according to claim 1, wherein the standardizing the representative distance between fragments of step iii) comprises:

1) setting a reference group including normal samples from which chromosomal aneuploidies are not detected;

2) calculating a mean (reference_mean) and standard deviation (reference_standard_deviation) of RepFDs for respective chromosomes observed in the reference group;

3) conducting Z standardization by applying the reference_mean and reference_standard_deviation calculated in step 2) to Equation 1 below; and

Equation 1:

$$Z_{chr} = ((RepFD_{chr} - Reference_Mean_{chr}) / Reference_Standard_Deviation_{chr}) + 5$$

4) dividing Z_{chr} for the chromosome of the reference group calculated in step 3) by Z_{chr} for the chromosome for which presence of aneuploidy is to be determined.

5. The method according to claim 1, wherein the representative FD (RepFD) comprises at least one selected from the group consisting of a sum, difference, product, mean, median, quantile, minimum, maximum, variance, standard deviation, median absolute deviation and coefficient of variance of the fragment distance (FD), a reciprocal thereof, and a combination thereof.

6. The method according to claim 1, wherein the artificial intelligence model is selected from the group consisting of a convolutional neural network (CNN), a deep neural network (DNN), a recurrent neural network (RNN), and an autoencoder.

7. The method according to claim 1, wherein, when the artificial intelligence model is CNN and learns binary classification, loss function is represented by Equation 2 below, and when

the artificial intelligence model is CNN and learns multi-class classification, loss function is represented by Equation 3 below.

Equation 2: Binary classification

$$\text{loss}(\text{model}(x), y) = -\frac{1}{n} \left[\sum_{i=1}^n (y_i \log(\text{model}(x_i)) + (1 - y_i) \log(1 - \text{model}(x_i))) \right]$$

Model (x_i) = Artificial intelligence model output in response to i^{th} input

y = Actual label value

n = Number of input data

Equation 3: Multi-class classification

$$\text{loss}(\text{model}(x), y) = -\frac{1}{n} \sum_{i=1}^n \left(\sum_{j=1}^c (y_{ij} \log(\text{model}(x_i)_j)) \right)$$

Model (x_i) _{j} = j^{th} artificial intelligence model output in response to i^{th} input

y = Actual label value

n = Number of input data

c = Number of classes

8. The method according to claim 1, wherein the cut-off value of step d) is 0.5, and when the resulting value is 0.5 or more, it is determined that there is a chromosomal abnormality.

9. A device for determining a chromosomal abnormality based on artificial intelligence, the device comprising:

a decoder configured to extract nucleic acids from a biological sample and decode sequence information;

an aligner configured to align the decoded sequence with a reference genome database;

a data generator configured to generate vectorized data using nucleic acid fragments based on aligned sequence information (reads), wherein the generated vectorized data is a

grand canyon plot (GC plot) or a genomic castle wall plot (GCW plot); and

a chromosomal abnormality determiner configured to input the generated vectorized data into a trained artificial intelligence model, analyze the data, and compare the resulting value with a cut-off value thereby to determine whether or not a chromosomal abnormality is present;

wherein the data generator is further configured to generate the GC plot by calculating a distribution of aligned nucleic acid fragments in each chromosome bin based on the distance between the nucleic acid fragments by:

- i) dividing chromosomes into predetermined bins;
- ii) calculating the distance between nucleic acid fragments (fragment distance, FD) aligned in each bin;
- iii) determining a representative distance between fragments (RepFD) of each bin based on the fragment distance calculated in each bin;
- iv) normalizing RepFD by dividing the representative distance between fragments (RepFD) calculated in step iii) by a representative total nucleic acid fragment distance; and
- v) generating a GC plot with the order of respective bins on an X-axis and the normalized value calculated in step iv) on a Y-axis; and

wherein the data generator is configured to create the GCW plot by:

- 1) calculating the distance between the aligned nucleic acid fragments (FD) for each chromosome;
- 2) determining a representative distance between fragments for each bin based on the distance calculated in step i);
- 3) standardizing the representative distance between fragments in each chromosome determined in step ii);
- 4) selecting a portion of chromosomes excluding chromosomes for which presence of aneuploidy is to be determined as control chromosomes; and
- 5) generating the GCW plot with the control chromosomes and the chromosomes for which presence of aneuploidy is to be determined sequentially and alternately aligned on the X axis and the standardized calculated values of the respective chromosomes on the Y axis,

wherein the sequentially and alternately alignment is aligning ([calculated value of step ii) of the control chromosome], [calculated value of step ii) of the chromosome for which presence of aneuploidy is to be determined]) along the x-axis and

wherein the artificial intelligence is trained to distinguish between vectorized data of normal chromosomes and vectorized data of abnormal chromosomes.

10. A computer-readable storage medium including an instruction configured to be executed by a processor for detecting a chromosomal abnormality through the following steps comprising:

a) obtaining sequence information using extracted nucleic acids from a biological sample;

b) aligning the obtained sequence information (reads) with a reference genome database;

c) generating vectorized data using nucleic acid fragments based on the aligned sequence information (reads); and

d) inputting the generated vectorized data into a trained artificial intelligence model, analyzing the data, and comparing the resulting value with a cut-off value to determine whether or not a chromosomal abnormality is present,

wherein the vectorized data of step (c) is a grand canyon plot (GC plot) or a genomic castle wall plot (GCW plot),

wherein the GC plot is generated by calculating a distribution of aligned nucleic acid fragments in each chromosome bin based on the distance between the nucleic acid fragments using a process including the following steps:

i) dividing chromosomes into predetermined bins;

ii) calculating the distance between nucleic acid fragments (fragment distance, FD) aligned in each bin;

iii) determining a representative distance between fragments (RepFD) of each bin based on the fragment distance calculated in each bin;

iv) normalizing RepFD by dividing the representative distance between fragments (RepFD) calculated in step iii) by a representative total nucleic acid fragment distance; and

v) generating the GC plot with the order of respective bins on an X-axis and the normalized value calculated in step iv) on a Y-axis;

wherein the GCW plot is created using a process including the following steps:

1) calculating the distance between the aligned nucleic acid fragments (FD) for each chromosome;

2) determining a representative distance between fragments for each bin based on the distance calculated in step i);

3) standardizing the representative distance between fragments in each chromosome determined in step ii);

4) selecting a portion of chromosomes excluding chromosomes for which presence of aneuploidy is to be determined as control chromosomes; and

5) generating the GCW plot with the control chromosomes and the chromosomes for which presence of aneuploidy is to be determined sequentially and alternately aligned on the X axis and the standardized calculated values of the respective chromosomes on the Y axis,

wherein the sequentially and alternately alignment is aligning ([calculated value of step ii) of the control chromosome], [calculated value of step ii) of the chromosome for which presence of aneuploidy is to be determined]) along the x-axis and

wherein the artificial intelligence model of step (d) is trained to distinguish between vectorized data of normal chromosomes and vectorized data of abnormal chromosomes.